## Metyis | LinkedInMetyis Data Engineer Technical Assignment

### Project name: Creating Sales ETL

### Author: Konrad Wroński

### Project Description

This project invloves creating an ETL pipeline using PySpark. The process should read all of the files, merge them and write them to *cleansed* folder. The files are representing sales of a retail company. On top of data loading and processing, exploratory data analysis was performed.

### Approach

1. Creating PySpark session
2. Fixing column naming – lowercase and removing whitespaces
3. Creating schema and empty data frame to store merged files
4. Custom *load* function to process the csv files
5. *For loop* to interatively load and merge files
6. Handling missing values – removal of the rows, as missing values occur in multiple columns simultaneusly
7. Deduplication of the data using *Window function*
8. Data transformation – spliting the *Order Date* column for easier analysis
9. Loading the data into *cleansed* folder as parquet files with month and year partition
10. Data quality check – asserting that no data loss occurred with *assert.*

Parquet file has been chosen as a destination format as it is: columnar and supports compression

Partitioning strategy for this file is per month and year. The reason for this choice is that sales data are often used to compare certain time frames.

### Data Exploration

### A graph showing the number of products AI-generated content may be incorrect.

### A graph with a line AI-generated content may be incorrect.

### A graph of blue rectangular bars AI-generated content may be incorrect.